Comprehensible video acquisition for caregiving scenes—How multimedia can support caregiver training

AUTHOR(S):

Nakamura, Yuichi; Kondo, Kazuaki; Mashimo, Taiki; Matsuoka, Yoshiaki; Ohtsuka, Tomotake

Comprehensible Video Acquisition for Caregiving Scenes — How multimedia can support caregiver training

Yuichi Nakamura[1], Kazuaki Kondo[1], Taiki Mashimo[1], Yoshiaki Matsuoka[2], and Tomotake Ohtsuka[3]

[1] ACCMS, Kyoto University, Sakyo, Kyoto, Japan,
{yuichi,kondo}@media.kyoto-u.ac.jp
[2] Faculty of Health Science, Aino University, Ibaraki, Osaka, Japan
[3] Nishikagawa Hospital, Mitoyo, Kagawa, Japan

Abstract. Caregiving requires a variety of techniques that directly affect care-receivers' quality of life. However, caregivers have difficulties in objectively observing and confirming their skills and whether they are providing care-receivers with appropriate and acceptable care. Videotaping is often used in caregiver training for this purpose. This paper introduces a novel method for enhancing such videotaping using information media technology, capturing video using multiple cameras and editing the videos automatically into a single-stream video to achieve a recording purpose. The purpose is to provide an effective means of observing and learning caregiving from typical important points of view. Our experiments showed that videos obtained through our proposed method are effective for the intended purposes.

1 Introduction

Many kinds of interaction between caregivers and care-receivers occur in caregiving everyday. One of the most important problems for caregiver is to make their interactions smooth and comprehensible to care-receivers, because they directly affects care-receivers' the quality of life (QOL). Caregivers need specific skills and training in communication to enable care-receivers to understand their intentions fully. Videotaping has been used often for training. Caregivers can testing and improve skills that they have difficulty evaluating on their own. They can also learn how experts behave and pay attention in caregiving.

While videotaping is a promising method for learning and training, it is tiresome, and skills must be recorded and paid attention to are often difficult to capture precisely. For this purpose, we propose video acquisition support in caregiving scenes by introducing recent media technologies, such as smart video capturing using multiple cameras and automatic editing of captured videos. Those techniques provide not only comprehensible videos but also different means of reviewing videos corresponding to different viewpoints.

One important point that we need to focus on is care-receiver's perception, *e.g.*, how a care-receiver perceives a caregiver's approaches, how a care-receiver

pays attention, how a care-receiver understands care. This viewpoint enables us to understand what kinds of reaction caregiving methods might cause, and why they occur. Another possible point is an experts' skill, *e.g.*, to which portion experts pay attention most and how they make their intentions understood by care-receivers. We can also consider other viewpoints, for example, the perspectives on usability of equipment in caregiving settings.

With the help of video captured considering the above viewpoints, we can expect that caregivers and care-receivers' families have favorable opportunities for improving their skills with better understanding of care-receivers' characteristics. This leads to better care-receivers' QOL. In this paper, we demonstrate the potential of our scheme through preliminary experiments involving simple tasks that appear commonly in daily care.

In the following sections, we first present background and related works in Section 2, the purpose and general framework for taking caregiving videos in Section 3, and possible video capturing and editing techniques in Section 4. We describe our preliminary experiments showing the potential of our scheme in Section 5.

2 Background and Related Work

This research aims to provide training supports for human-centered care and care-related knowledge to various communities. Humanitude[1, 2] is a powerful methodology for caregiving. Its key idea is to consider the perceptual and cognitive characteristics of care-receivers, and to make full use of human communication channels for ensuring sufficient information can be shared among care-receivers and caregivers. Methods and skills for eye contact, touch, speech, and standing have been intensively discussed and put into practice. Practice of this methodology drastically reduces care-receivers' undesirable and often aggressive behaviors, improves caregivers' satisfaction, and accordingly improves care-receivers' QOL.

Based on this background, the idea arose that one effective way for supporting caregiving is to help caregivers understand care-receivers' nature by showing how the caregiving method changes care-receivers' perceptions and reactions. Recent devices and media technologies on capturing and editing videos have potential for supporting the above purpose. Topics on video acquisition and handling, *e.g.*, automated video capturing, automated video editing, and video indexing and retrieval have been intensively explored, and their possibilities have been demonstrated. The application most related to the above purpose is automated lecture archiving and video content production[3–5]: automated cameras shoot at people or objects, and the system recognizes humans, objects, and events, which are then used for automated editing and summarizing [6].

In addition, wearable camera devices have been developed for a variety of purposes, such as Lifelog, daily healthcare, and remote medicine[7, 8]. First person vision (FPV), that is, taking videos or pictures with a small camera attached to the head or body, is closely related to our topic. We can record what we see

(a) catch sight (b) talk to (c) touch (d) explanation

Fig. 1. Typical shots focusing on actions in interactions

and what we experience, and we can review the record on our demand, *e.g.*, recalling daily events or looking for a lost object[9, 10]. We can expect that such technologies will provide substantial assistance in recording and utilizing caregiving scenes and enable us to understand how to meet care-receivers' demands.

3 Purpose of capturing caregiving scenes

3.1 What needs to be captured

An essential problem is how we can notice and understand important points, when we are learning or practicing caregiving. For example, we need to know how a care-receiver perceives a caregiver's approach and touch and understand how their undesirable reactions such as Behavioral and Psychological Symptoms of Dementia (BPSD) are sometimes caused by caregiving. Those are tightly related to interactions between caregivers and care-receivers, spatial relationships and physical touches, and the facilities or environments in a care setting. Let us consider those aspects in the following.

Actions in interactions: Interactions between caregivers and care-receivers occur through several channels, including eye contact, facial expression, speech, and touch. It is important to see details in what signals and movements are taken in which ways, and to which portion attention has been paid. Figure 1 shows typical scene captures that are focused on the above points. Figures 1 (a–d) show how the caregiver entered the sight of the care-receiver, initiated the interaction by talking to the care-receiver, touched the care-receiver on the leg, and conducted descriptive conversations, *i.e.*, explained what the person was about to do, respectively.

Attention and perception: Perception and attention, *i.e.*, how people pay attention to each other, what is perceived, and how it is recognized provides essential information in caregiving. Knowing care-receivers' perception and attention is particularly important, since caregivers must exercise considerable skill and effort to enable care-receivers with reduced sensory efficiency to understand their intentions. Figure2 shows scene captures that explain perception and attention well. Figure2(a) and (b) show where or to which object a care-receiver paid

(a) attention of a care-receiver (b) target of attention (c) not noticed by care-receiver (d) attention of a caregiver

Fig. 2. Typical shots focusing on attention and perception

(a) overview (b) posture (c) relationship to an object (d) tools and furniture

Fig. 3. Typical shots focusing on spatial relationships and environments

attention. Figure2(c) shows an undesirable example in which the care-receiver did not notice the caregiver's action, for which Figures 1 (a) and (b) already showed an improved example in which the caregiver successfully caught the eye of the care-receiver. Figure 2(d) shows where and what the caregiver paid attention to. Joint attention such as one person directing the other's attention is also an important factor.

Position and environment: The spatial and positional relationship of a caregiver and a care-receiver is also an important factor, and how facilities are used or affect the manner of caregiving is also useful information. Figure 3 shows typical scene captures relating to the above information. Figure 3(a-c) show an overview of the scene, a posture of a caregiver and a care-receiver, the spatial relationship to the objects, and how tools and furniture were arranged, respectively.

3.2 How scenes need to be presented

An audience naturally attempts to find continuity and causality among neighboring shots in a video. An appropriate combination of shots provides rich clues of cause-and-effect relationships, as well as specific information in each shot.

For example, a caregiver's skill at catching the sight of a care-receiver would be explained well by a combination of the shots of the caregiver's action of coming closer and getting into the view of a care-receiver, the action of touching the care-receiver, and the care-receiver's and the caregiver's faces, together with a subjective view of the caregiver from the care-receiver's viewpoint.

Typically, all shots cannot be packed into one video stream, because multiple cameras provide views of different targets simultaneously. Replaying all video streams in parallel also fails to provides a comprehensible explanation of a scene. Watching multiple videos in parallel is usually tiring, and it is difficult to pay attention to the correct portion[4]. For example, we have difficulty in watching both the caregiver's attention and the care-receiver's attention simultaneously. Consequently, we need video editing to present such scenes comprehensibly by choosing an appropriate shot at each moment. This kind of problem has been discussed well as "editing" in film studies[11], where we can find useful knowledge and techniques in actual movies and research.

The problem is how to invoke such knowledge and techniques to satisfy our purposes, some of which involve the following.

Overview of caregiving scene: This focuses primarily on how care actions and events occur in a scene, and what results are obtained.

How a care-receiver receives care: This focuses primarily on how a care-receiver perceives and understands a caregiver's actions, and on how a care-receiver accepts, misses, or rejects it.

How a caregiver applies skills for care: This focuses primarily on how a trainee or an expert approaches a care-receiver, what he or she pays attention to, and how he or she performs care actions.

Although those purposes are not always mutually exclusive, they sometimes conflict because it is very difficult to pay attention to different targets simultaneously.

4 Utilizing knowledge of film studies and media technology

Film studies have proved that audience perception and understanding heavily depend on shots and editing schemas.

4.1 Category of shots

Let us first consider shots. In film studies, shots are categorized based on target size in a screen *e.g.*, close-up, bust, or medium shot. Another categorization of shots is based on viewing position and angle, *e.g.*, point-of-view (POV) shot[5], bird's-eye view, mobile-view shot[6].

We need to choose appropriate shots for typical purposes. Figure 4 shows typical shots.

[4] We are assuming that people without professional skills are analyzing videos. For a person with video analytics skills, multiple views with complete information could be the most powerful tools.

[5] A camera is shooting at the scene simulating the persons sight from the position of his/her eyes.

[6] A camera is moving, typically dollying.

(a) close-up (b) POV shot (c) medium shot (d) long shot

Fig. 4. Example of typical shot categories in Film Studies

- A close-up shot, as shown in Figure 4 (a), or a bust shot is preferable for showing facial expression and emotion of a care-receiver or a caregiver.
- A close-up shot, as shown in Figure 4 (b), is most suitable for showing the target to which attention is paid. A POV shot can be an alternative. Other shots, such as a medium shot (c) or a bird's-eye view shot, would partially explain attention.
- Presenting touch or physical interaction between a care-receiver and a caregiver requires the same types of shot as the above.
- Explanation of spatial relationships between a care-receiver and a caregiver requires longer and wider shots such as (c) or (d). A bird's-eye view shot is also acceptable. Those shots play a role of an "establishing shot" that provides overview of a scene and its environment.
- Presenting postures and movements of a care-receiver or a caregiver requires the same types of shot as the above, though a combination of medium or closer shots could be alternatives.

We assume that those shots and their combinations have potential to explain caregiving scenes. What we need is to set up a camera for taking each of the above types of shots. Sufficiently many cameras must be located at appropriate locations with appropriate focal lengths.

4.2 Editing technique

As discussed above, we consider view switching among multiple cameras as an editing scheme. To simplify this process, we use the automated editing scheme proposed by Ogata et al.[6]. In that scheme, video editing is considered as the problem of assigning an appropriate shot to each video unit segment. Every possible pattern of assignment is listed and scored using evaluation functions, and the best pattern is chosen. Appendix A presents the formal description of this editing.

Figure 5 shows the flow of computation, which consists roughly of three steps, *pre-scoring*, *candidate searching*, and *post-scoring and selection*. The correspondence between the editing scheme and our purpose in videotaping can be explained briefly as follows.

In the pre-scoring step, the relevance of each shot at each time is evaluated based on the matching between characteristics of the shot and the purpose of

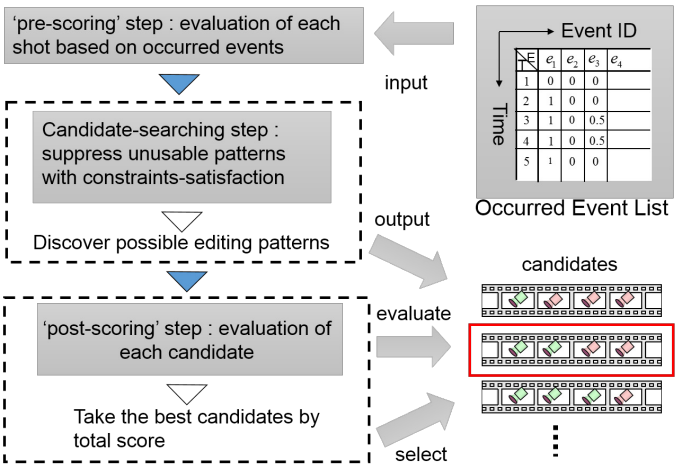


Fig. 5. Flow of computation

videotaping. It is related to events occurring in the scene. Table 1 shows the rough idea. Basically, we assign better scores to the shots that match an editing purpose, *i.e.*, the shots that should be paid attention for the editing purpose. For example, when a caregiver is talking to a care-receiver, the care-receiver's POV shot is assigned a high score if the purpose of video is to focus on the care-receiver's perception.

In the candidate-searching step, possible editing patterns that satisfy constraints are searched using a constraint programming. Some constraints are derived from common and conventional editing schemas in film studies, and others from the nature of human perception. For example, "prohibiting too frequent shot changes" is a typical rule for avoiding dizzying patterns that are difficult to understand. A substantial number of unnecessary editing patterns are drastically eliminated at this step.

In the post-scoring step, each editing pattern receives a final score. This score evaluates primarily the appropriateness of combinations of consecutive shots that are not fully fixed in the pre-scoring step. For example, a POV shot is preferable if it has a preceding close-up shot of the viewing person, because audience sometimes have difficulty in understanding whose POV it is.

After scoring is completed, the editing pattern, *i.e.*, the shot sequence, that received the highest score, is selected as the result.

5 Preliminary Experiments

Objective: The objective of the experiments is to confirm that videos obtained using our scheme are comprehensible and help audiences notice the focused points. We chose two typical caregiving scenes, serving meal to a care-receiver

Table 1. Example of editing rules (how each shot fits each purpose. Values range between 0 (no match) and 1 (best match)).

Editing purposes (what is focused)
(a) how the caregiver and the care-receiver behave (overview).
(b) how the care-receiver receives care (care-receiver's attention and reaction).
(c) how the caregiver gives care (caregiver's attention and care-giving skill).

shot	(a)	(b)	(c)
caregiver talks to care-receiver			
long shot	0.8	0.3	0.5
close-up of caregiver	0.3	0.3	0.7
....	...	...	...
POV of care-receiver	0.3	1.0	0.7
caregiver serves something			
long shot	0.7	0.2	0.3
medium shot of caregiver	0.3	0.3	0.9
....	...	...	...
close-up of care-receiver	0.5	0.5	0.7
POV of care-receiver	0.7	1.0	0.5

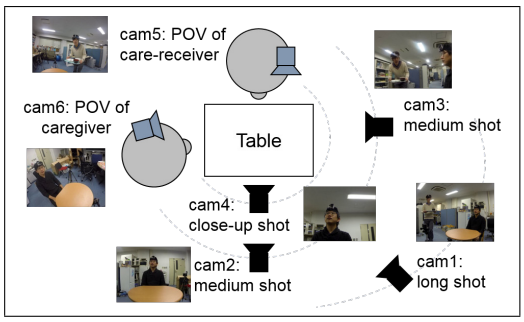


Fig. 6. Camera arrangement

Fig. 7. Multiple videos: all videos are presented and replayed simultaneously in a tiled format

and helping a care-receiver stand up. These two scenes frequently appear in daily caregiving, in which problems sometimes occur as a result of a care-receiver's reduced sight, attention, or memory capacity. We simulated the above two types of care scenes by some of the authors.

Shots and events: Two head-mounted cameras and four fixed cameras were used, as shown in Figure 6. Head-mounted cameras are usually allowable in training, and are allowable if care-receivers are willing to cooperate to improve caregiving. However, the location of a fixed camera is often limited because of the spatial arrangements of a caregiving venue. We assumed the behaviors of the care-receiver and the caregiver are observed sufficiently as events, *e.g.*, speaking, walking, or touching. For each recorded video, we manually detected those events and annotated them with time of occurrence. Automatic detection of events is left for future work.

Editing and purposes: We compared three purposes of videotaping: (a) focusing on how a caregiver and a care-receiver behaved and felt, (b) whether

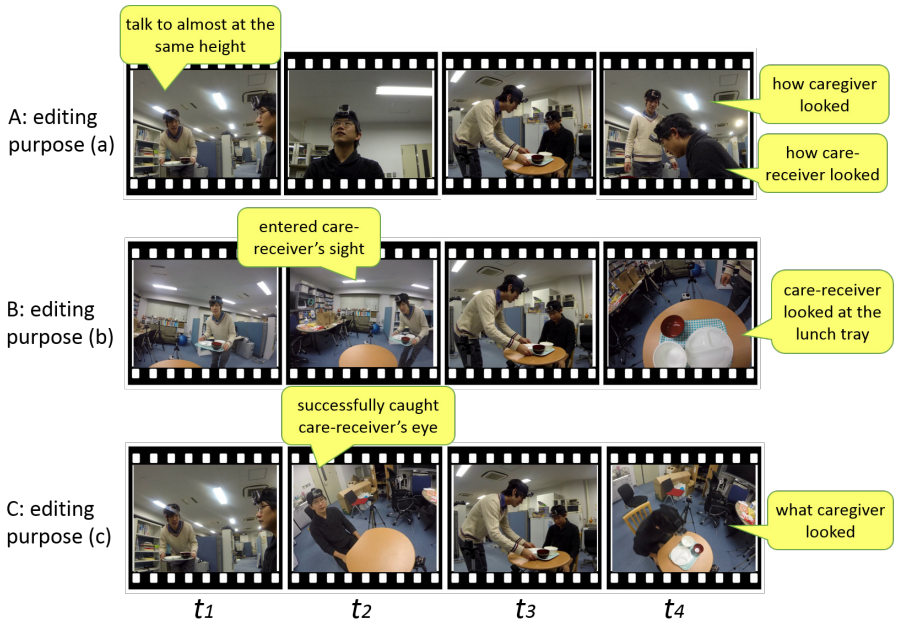


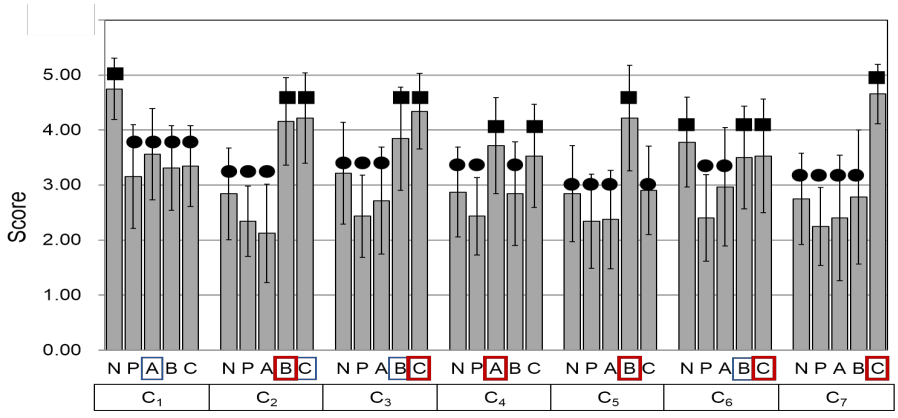
Fig. 8. Edited movies

care was perceived sufficiently by the care-receiver, and (c) a caregiver's attention and skills. Table 1 contains an example of rules for each videotaping purpose.

Results: Figure 8 presents the obtained movies in a film strip view. Comparing (a), (b), and (c), we can see differences among the movies. Movie (a) presents an overview of how both persons behaved and felt. Movie (b) emphasizes the care-receiver perception: (b) at t_2 and t_3 show what the care-receiver looked when the caregiver talked to him, and (b) at t_4 indicates that the care-receiver probably understood that the lunch tray was served to the care-receiver. Movie (c), in contrast, emphasizes where the caregiver looked and how the caregiver behaved. Movie (c) at t_2 shows that the caregiver looked at the care-receiver's eye and confirmed eye contact, and movie (c) at t_4 shows how the caregiver checked the care-receiver's recognition.

To verify the above observations quantitatively, we asked eight participants to provide subjective evaluations, scoring each video according to the criteria shown in Figure 9. For comparison, we added two types of video, a long shot without editing and a combination of all views replayed simultaneously, as shown in Figure 7. The participants watched those movies and assigned scores based on whether the specified information is difficult or easy to discern from the movies.

The graph in Figure 9 shows the results. Score 5 is the most positive (easy) and 1 is the most negative (difficult). The graph shows clear differences among videos. For each criterion, the difference between group "■ (suitable)" and group "● (not suitable)" is statistically significant at the 1% level.



Videos
N: Long shot with no editing
P: Multiple videos are presented in tiled format and replayed simultaneously
A: Editing purpose (a)
B: Editing purpose (b)
C: Editing purpose (c)

Criteria
C_1: How care is given in the scene.
C_2: Degree to which eye contact is established and maintained.
C_3: The the caregiver's position is appropriate when the caregiver talked to the care-receiver.
C_4: How the care-receiver feels
C_5: How the care-receiver perceives or recognizes the care
C_6: How the caregiver touches the care-receiver
C_7: Which object or place the caregiver looks at

Fig. 9. Subjective evaluation result

Notice that the simultaneous replay of all views in a tiled arrangement (P) obtained almost the worst score for any criterion, strongly supporting the necessity of editing. Long shot without editing (N) obtained good scores for C_1 and C_6, which request primarily overview and spatial information. Next, let us check how the obtained videos meet the editing purposes. The rectangles with thick lines enclosing P, N, or A-C below the graph show videos that are primarily related to the criteria, and the rectangles with thin lines show marginally related videos. Editing (a) obtained good scores for C_4, which is the primary purpose of editing (a); however, an overview of the scene (C_1) is not well indicated. Editing (b) obtained almost the best scores for C_2, C_5, and C_6, which are related to how the care-receiver perceived and received care. Editing (c) obtained almost the best scores for C_2, C_3, and C_7, which are related primarily to the caregiver's skills.

Discussion: The result supports our idea that appropriate video capturing and editing facilitates better understanding of care scenes. However, the participants for verification were not skilled, which can be considered to be the case of novice

trainees or patients' family members. Systematic evaluation including skilled professionals is also necessary for future work.

Moreover, we still need further work to ensure that our scheme will be applicable to actual caregiving scenes. More systematic evaluations for various situations are necessary. Image processing and speech recognition must be investigated to detect events, such as speech, touch, and important motions. Such automated event detection and design would substantially reduce the time and cost of actual practice. We also need a complete system design, including camera arrangement, for capturing sufficient information.

6 Conclusion

This paper introduced a novel scheme of video content acquisition and editing for caregiving scenes. The scheme utilizes knowledge of film studies and media technology for obtaining appropriate videos for typical purposes. The preliminary results are convincing in that the obtained video emphasizes the purpose of videos, such as understanding care-receiver's perception and caregiver's attention, and assists an audience in noticing important features.

We have substantial room for future work. Further experiments with a variety of caregiving scenes and video purposes are necessary, as is verification in actual training. In addition, we need to put further efforts into automating event detection and possibly capturing scenes more actively, *e.g.*, pan/tilt or other camera movements, since this scheme aims to reduce the tiresome work of videotaping.

References

1. Honda, M., Gineste, Y., Marescotti, R.: Introduction to Humanitude (in Japanese), Igaku Shoin (2014)
2. Phaneuf, M.: "The concept of humanitude as applied to general nursing care", `http://www.infiressources.ca/fer/depotdocument_anglais/↩`
`↪the_concept_of_humanitude_as_applied_to_general_nursing_care.pdf`
(accessed April 15, 2015)
3. Atarashi, Y. et al.: Controlling a Camera with Minimized Camera Motion Changes under the Constraint of a Planned Camera-work, Proceedings of International Workshop on Pattern Recognition and Understanding for Visual Information Media 2002, pp.9-14 (2002)
4. Ozeki, M., Nakamura, Y.: Evaluation of Self-Editing Based on Behaviors-for-Attention for Desktop Manipulation Videos, Proc. IEEE Int'l Conference on Multimedia and Expo, MA2-L5.2 (2006)
5. Yamaguchi, S., Ohnishi, Y., Nishino, K.: The Design of an Automatic Lecture Archiving System Offering Video Based on Teacher's Demands, Intelligent Interactive Multimedia: Systems and Services Smart Innovation, Systems and Technologies, Volume 14, pp 599-608 (2012)
6. Ogata, R., Nakamura, Y., Ohta, Y.: Computational Video Editing Model based on Optimization with Constraint-Satisfaction, Proc. Fourth Pacific-Rim Conference on Multimedia, CD-ROM, 2A1-2 (2003)

7. Garner, P., Collins, M., Webster, S., Rose, D.: The application of telepresence in medicine, BT Technolo J, Vol.15, No.4, pp.181–187 (1997)
8. Gemmell, J., Bell, G., Luede, R.: MyLifeBits: a personal database for everything, Communications of the ACM, vol. 49, Issue 1, pp. 88–95 (2006)
9. Hodges, S. et al.: SenseCam: A Retrospective Memory Aid, Ubicomp 2006, LNCS 4206, pp. 177-193 (2006)
10. Berry, E., et al.: The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report, Neuropsychological Rehabilitation, Vol.17 No.4/5, pp.582-601 (2007)
11. Bordwell, D., Thompson, K.: Film Art: An Introduction, 5th edition, McGraw-Hill (1997)

A Definition of computational editing

In the following, we briefly introduce the editing scheme used in our method. Please see [6] for details. The definition is modified for clarity, but the computational scheme is the same.

The computational model of editing is composed formally of five elements:

$$\text{Editing} = \{S, V, E, O, C\} \tag{1}$$

The explanation of the above terms is as follows:

Shots (S): S is a set of shots *i.e.*, $S = \{s_0, \ldots, s_n\}$, where s_i is a shot, *e.g.*, "a bust shot of person A," "a long shot of person B,".

Shot assignment (V): V is a sequence of shot assignments to video segments, each of which has a length, *e.g.*, 0.5 or 1 second. One shot in S is assigned to each video segment, *i.e.*, $V = \{s_i, \ldots, s_k\}$,

Events (E): E is a collection of events $\{e_i\}$ occurring in the scene, for example, "person A spoke," "person B laughed," or "person A touched person B." If e_i occurs at time t with a certainty of 0.9, we denote it as $e_i(t) = 0.9$.

Evaluation (O): O is a collection of objective functions $\{o_i(V, t)\}$, each of which assigns a score for a shot assignment at time t. The criterion can be comprehensibility, entertainment quality, or one of many other factors.

Constraints (C): C is a set of constraints. Some of the editing rules are constraints, *e.g.* "do not use shots longer than t_n seconds." The number of candidates can be reduced by applying the constraints.

The objective of this model is the optimization of G in the following formula.

$$G = \sum_{t=0}^{t_{max}} \sum_{i=0}^{i_{max}} o_i(V, t) \tag{2}$$

In other words, the objective is to find the best assignment of shots to video segments that maximizes evaluation value G based on O and satisfies constraints C. For this purpose, we first use a constraint programming to obtain candidates and select the best-scored editing